

DOCUMENT RESUME

ED 282 903

TM 870 301

AUTHOR Garrido, Mariquita; Payne, David A.
TITLE An Experimental Study of the Effect of Judges' Knowledge of Item Data on Two Forms of the Angoff Standard Setting Method.
PUB DATE Apr 87
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, DC, April 21-23, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; Comparative Analysis; *Criterion Referenced Tests; *Cutting Scores; Difficulty Level; Higher Education; *Interrater Reliability; *Judges; Minimum Competency Testing; Multiple Choice Tests; Statistical Distributions; Statistics; Test Items; Test Theory
IDENTIFIERS *Angoff Methods; *Standard Setting

ABSTRACT

Minimum competency cut-off scores on a statistics exam were estimated under four conditions: the Angoff judging method with item data (n=20), and without data available (n=19); and the Modified Angoff method with (n=19), and without (n=19) item data available to judges. The Angoff method required free response percentage estimates (0-100) percent, whereas the Modified Angoff method used a seven point scale (5, 20, 40, 60, 75, 90, 95). Item data were in the form of frequency distributions of multiple-choice exam option selections for five student ability levels. Results indicated significantly: (1) greater intra-judge variability for the Angoff method; and (2) higher minimum competency cut-off scores for judges with performance data available for reference than for judges who did not have data available. Acceptable levels of inter-judge reliability, judge internal consistency, and item internal consistency were also observed. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED282903

AN EXPERIMENTAL STUDY OF THE EFFECT OF JUDGES' KNOWLEDGE OF ITEM DATA
ON TWO FORMS OF THE ANGOFF STANDARD SETTING METHOD

Mariquita Garrido and David A. Payne
University of Georgia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as
received from the person or organization
originating it.

☒ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

D. A. Payne

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

ABSTRACT

Minimum competency cut-off scores on a statistics exam were estimated under four conditions: Angoff judging method with item data (n=20), and without data available, (n=19), and Modified Angoff method with (n=19), and without (n=19) item data available to judges. The Angoff method required free response percentage estimates (0%-100%) whereas the Modified Angoff method used a seven point scale (5, 20, 40, 60, 75, 90, 95). Item data were in the form of frequency distributions of multiple-choice exam option selections for five student ability levels. Results indicated significantly (a) greater intra-judge variability for the Angoff method, and (b) higher minimum competency cut-off scores for judges with performance data available for reference than for judges who did not have data available. Acceptable levels of inter-judge reliability, judge internal consistency and item internal consistency were also observed.

AN EXPERIMENTAL STUDY OF THE EFFECT OF JUDGES' KNOWLEDGE
OF ITEM DATA ON TWO FORMS OF THE ANGOFF STANDARD SETTING METHOD

Mariquita Garrido and David A. Payne
University of Georgia

The widespread development and application of criterion-referenced (minimum competency) tests has highlighted many political and technical problems in test use. With the increase in selection, placement, certification, promotion, and graduation decisions being made on the basis of data derived from these kinds of tests come questions about standard setting (Hambleton and Eignar, 1980).

Berk (1986) has identified over 38 different methods that have been developed to assist in either setting standards or to adjust already existing data to yield standards. Basically two kinds of approaches are used. The first relies on the absolute but subjective estimates of judges and the second relies either directly or indirectly on an empirical data base to aid in decision-making. One of the frequently used purely judgmental methods is that proposed by Angoff (1971). This method requires judges to examine each test item and estimate the probability that a minimally competent subject would answer the item correctly. This probability translates to a proportion of a sample who would pass the item or respond in a particular direction. A modification of Angoff's procedure narrows the response scale to seven points (5, 20, 40, 60, 75, 90, 95) as it was found that judges often had difficulty in using the expanded scale (ETS, 1976). Berk (1986) concludes that the Angoff approach shows the best balance in terms of technical adequacy and practicality.

A procedure that is sometimes used to enhance the technical adequacy of a judging method is to provide performance data to the judges. Cross, Impara, Frary, and Jaeger (1984) found that initial standards changed after feedback of normative data was provided to judges. The effect of the information was to lower standard cut-off scores for the National Teacher Examinations.

The present authors did not find studies where the Angoff (1971) and Modified Angoff (1976) procedures have been compared, or where the effect of knowledge of item performance data on Modified Angoff

estimated cut-off scores had been assessed. It was to fill this gap in the research that the present study was undertaken.

METHOD

Instrumentation

Items to be judged were selected from an existing midterm exam of 30 items which had been administered to 55 examinees and had an internal consistency reliability index of .77. Twenty items were selected using an 18 cell table of specifications to assist in maintaining content balance. Topics included central tendency, variability, notation, graphing and standard scores. Items with difficulty indices ranging from .51 to .98 were used for the reference data.

Four booklets were prepared, one for each method/condition combination, i.e. Angoff with and without reference data, and Modified Angoff with and without data. Each booklet contained two pages of introduction and instructions to the judges. In addition the booklets with reference data contained an illustration of how to interpret the item analysis data. The correct answers for all items were indicated and in addition for the data condition groups item data in the form of difficulty indices, and the frequency distributions choices of options by five levels of ability were presented. All responses were made to the following general query

What percentage of your minimally competent group of students do you judge will correctly respond to this item?

Minimally competent was defined in the judging directions as being a "C" level student. For the With Data condition the item performance information was presented concurrently with the judging task. No training of judges was involved. For the Angoff method a free response scale ranging from .00 to 1.00 could be used, whereas the Modified Angoff judging group was required to use one of the following seven fixed percentages 5%, 20%, 40%, 60%, 75%, 90%, and 95%. Demographic data were also gathered on the judges.

Judging Sample

Judges used for this study were graduate students who had completed at least an introductory course in educational statistical methods. Almost half the group had also completed course work in research design, computer programming, intermediate statistics or evaluation methods. The total voluntary sample of 78 judges had a mean age of 32.8 years and a mean teaching experience of 5.1 years. There were 24 males and 54 females.

Data Collection

The four booklets were randomly assigned to judges yeilding a 2x2 design and the following cell sizes:

		METHOD	
		<u>Angoff</u>	<u>Modified Angoff</u>
CONDITION	With Data	20	19
	Without Data	19	19

All judges operated independently. The criterion data collected from each judge was his/her estimate of the percentage of a group of minimally competent students that would respond correctly to the item. The cut-off score for each of the four groups was the average of the judges' estimated cut-off score in that group.

RESULTS

Basic descriptive data are presented in Table 1. Judges estimated cut-off scores tended to be higher for those judges given item performance data than for judges not given such data. In addition variability was greater for the Angoff method than the Modified Angoff method as one would expect due to the restricted response range imposed by having only seven possible percentages to choose from under the Modifed Angoff condition. All estimated item difficulty distributions evidenced some degree of negative skewness.

An index of within judge homogeniety was created in order to study the variability of each judge's responses over the 20 items. This index was the standard deviation for each judge using his/her own

mean as the reference point. These indices are included in Table 1. Results of a two-way analysis of variance revealed a significant between judging methods (Angoff = .19, Modified Angoff = .21) ($F = 4.70$; $df = 1,73$; .01). Although statistical significance was found here, it is doubtful that there are any practical implications of the finding. No significance between data condition differences, nor significant interaction were noted.

Differences Between Judging Methods

A two-way analysis of variance did not reveal a significant difference between the mean estimated cut-off score for the Angoff method ($X = 12.03$) and the Modified Angoff method ($X = 12.14$) ($F = .09$ $df = 1,73$). No significant method by condition interaction was noted ($F = .18$, $df = 1,73$).

Effects of Item Information

It was found that the With Data group mean of 12.96 was significantly different from the Without Data mean of 11.19 ($F = 14.56$, $df = 1,73$, $p < .01$).

A supplementary analysis was run to calculate the extent of the effect of presentation of item performance data (by using the item difficulty indexes) had on the judges' item responses. For each group, the mean responses for each 20 items were determined and then correlated against the 20 item difficulty indexes. The Pearson product moment correlation coefficients for the four groups were as follows:

<u>Method/Condition</u>	<u>r</u>
Angoff/with performance data	.98
Angoff/without performance data	.72
Modified Angoff with performance data	.98
Modified Angoff without performance data	.61

Using Fisher's Z transformation of the r coefficients, six pair-wise tests for differences among the independent correlation coefficients were computed. It was found that the correlation coefficients do not differ significantly from each other when judges were under the same item performance data condition (either with access to item performance data condition or without access to item performance data condition). However, comparisons across item performance data condition, regardless of judging process used, showed

the correlation coefficients were sufficiently different from each other. Thus, the test of independent correlation coefficients showed significantly higher correlations between different presentations of performance data conditions but within the same performance data conditions. These data supported the previously stated results indicating that differences in cut-off scores did occur within a particular method between judges receiving item performance data and judges not receiving item performance data.

Figure 1 displays a plot of the item difficulty index and the mean item probability for the 20 test items used in the study. As seen from the plot of group means for each item, the item means for the groups of judges presented with the item performance data closely followed the line flow of the item difficulty index. However, the item means for the groups of judges not presented with item performance data also followed the curve of the plots of the item difficulty index data. The figure suggests that when presented with item performance data, judges did use the item difficulty index in setting their item probabilities. However, since the item means for judges who were not presented with the item performance data still followed the contour of the item difficulty index plot, it could be suggested that judges used in this study were able to distinguish item difficulty in setting their estimated cut-off scores.

Inter-judge Reliability

The inter-judge reliability coefficients were computed using Ebel's variance components model (Ebel, 1951). The inter-judge coefficients are summarized in Table 1. Having access to the item performance data tended to enhance inter-judge reliability as the two largest coefficients (.94 and .96) occurred under the With Data condition, and the smallest (.74 and .79) under the Without Data condition. The same pattern is observed when the coefficients are calculated for a single judge.

Judge Internal Consistency

Cronbach's Alpha (Cronbach, 1951) was used to index the internal consistency of each judges' responses over the twenty item test. These coefficients are summarized in Table 1. All judges evidenced a high level of internal consistency.

As a secondary analysis, Cronbach's Alphas were obtained for test items on the judges within each group. This index was used to describe how well the items held together over the judges. The relevant data are again presented in Table 1. The coefficients are in an acceptable range, but with the Modified Angoff indices lower, probably due to the restricted response range.

No meaningful correlations were found between estimated cut-off scores and judges major area of study, years of teaching experience, number of statistics courses taken, sex, or age. This supports a finding by Behuniak, Archambault, and Gable (1982) of no significant relationships between judges demographic characteristics and the cut-off scores they set.

DISCUSSION AND CONCLUSIONS

The negatively skewed distributions of judges' estimated cut-off scores indicated a tendency for judges to be lenient in their estimation of the number of minimally competent students answering the test items correctly. In turn, this tendency led to higher cut-off scores. In addition, the higher cut-off scores were set by those judges receiving item performance data under both Angoff and Modified Angoff methods. These descriptive results suggested that even though judges tended to set high item probabilities, the judges expectations still were lower than actual students' performance.

The estimated cut-off scores were more dispersed under the Angoff method than the Modified Angoff method. This suggested that the presentation of a specified probability scale to the judges (Modified Angoff method) may have had some influence on the judges when item probabilities were determined. However, it may be that the presentation of the probability scale again simply reflected the possible problem of restriction of range--a tendency for judges not to use the extreme end points (.5 and .95) of the scale in setting item probabilities. Thus, the seven point probability scale under the Modified Angoff method was reduced to a five point probability scale lessening even more the possible variability occurring between judges' estimated cut-off scores.

Data related to within-judge item variability indicated that item performance data presented to judges did lessen the possible variability within each judge when compared to judges using the same method but without item performance data. The test of whether differences in within-judge item variability occurred significantly between methods or item performance data conditions did indicate significant variability between methods. Judges using the Modified Angoff method had slightly greater variability in within judge item variability than the judges under the Angoff method. This might imply that judges were more variable in their estimates of proportions when using the probability scale, and that in fact restriction of range may not have been a factor in the process.

Evaluation of the present data revealed no significant differences in the cut-off scores between the Angoff and Modified Angoff methods. However, the comparison between judges presented with item performance data and judges not presented with item performance data did reveal significant differences in judges' cut-off scores with the With Data condition yielding higher cut-off scores. This is contrary to the findings of Cross, Impara, Frary, and Jaeger (1984). In the Cross et. al. study however, the information was presented after the fact, and initial estimates were revised. It is hypothesized that in the present case any "leveling" took place concurrently with the judging. No significant differences were found in the interaction of method with presentation of item data. Overall, the external data presented with each item did influence the estimated cut-off scores established by the judges. The correlation between the item difficulty indices and the actual judges' item probabilities was to highlight the extent of the influence item performance data had on judges. An extremely high correlation was obtained for judges presented with item performance data. Although a high correlation was expected, the coefficient of .98 for these judges under the Angoff and Modified Angoff method almost questioned the creditability of the judges in their performance in the judging task. In other words, did judges truly understand the task of estimating the percentage of their minimally competent group of students answering the items correctly? Are judges able to conceptualize the performance of heterogenous group of students (as presented through the item difficulty indices) as

opposed to the performance of a homogeneous group of students? The test for differences among independent correlations further supported the statement that item performance data did influence the judges' responses in this judging task. The question remains, did the presentation of such information influence the judges to the extent that they disregarded their own judgements and relied solely on the item difficulty index in determining their probabilities? It should be noted that judges were told to use the item performance data in determining their item probabilities. Judges were not told how to use such information in their task. The results tended to indicate that judges in this study were not able to incorporate the item performance data with their past experience. It was the tendency of judges to take the item difficulty index at face value as the sole indication of success for a group of minimally competent examinees. Further use of the item performance data may require the standard setters to specify the exact steps judges will follow in the use of such external data through training sessions.

Between-judge reliabilities were computed under each condition. Judges presented with item performance data, regardless of method, had extremely high inter-judge reliabilities (.94 and .96). This high similarity between judges was moderated as one might expect when comparisons were made to single judges (.43 and .57). High inter-judge reliabilities were obtained for the two groups of judges not presented with item performance data (.74 and .79). However, when the reliabilities were stepped down to the single judge, wide variabilities were found for the single judge (.13 and .16). These results suggested that, as a group, judges were very similar in their judgements regardless of which method was used in the judgmental task. The wide range of variability evidenced with the single judge coefficient suggested the danger that could arise in setting cut-off scores with a single judge.

Based on Cronbach's alphas, judges were evaluated as being consistent in their responses to test items within each group. The judges given item performance data had the higher consistency coefficients.

The overall results suggested that groups of judges be used in setting cut-off scores. Judges not understanding the judging task,

and individual differences in judges' fatigue (more stringent on latter items) may be the cause of the wide variability in the single judge reliability index. The influence of a single judge was masked with the use of groups, and was considered to be an advantage. In conclusion, there was a positive effect with the use of item performance data in the standard setting judging task with high standards being set. The mean difference of 1.77 might not be practically significant. It would depend on how that difference translates to proportions of the tested population meeting or exceeding the standard. All four methods here have technical adequacy, are reliable, and have at least face credibility. Choice of method must be a function of nature of decisions and the, context of decision-making environment.

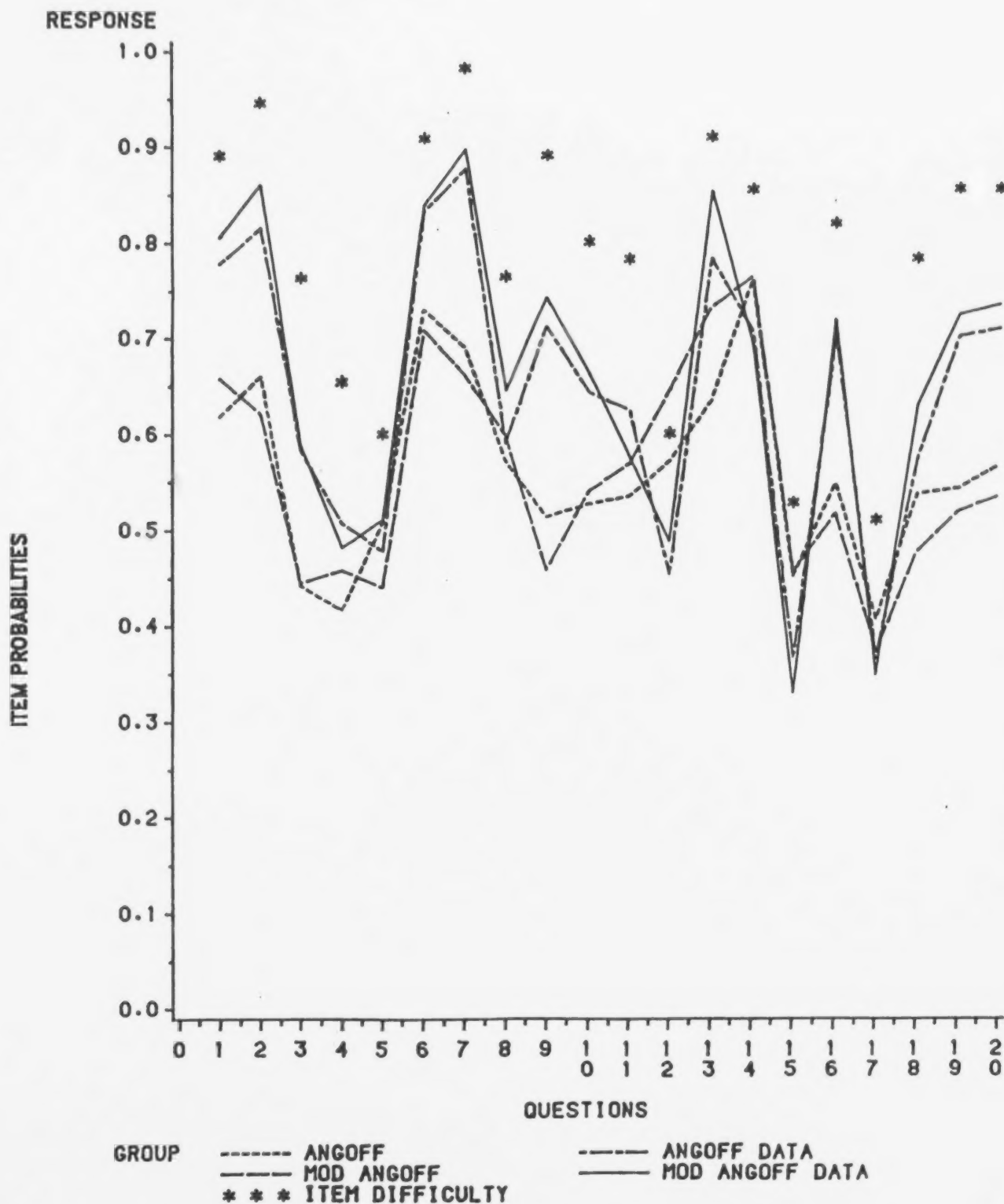
REFERENCES

- Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, D.C.: American Council on Education.
- Behuniak, P., Archambault, F.X. & Gable, R.K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 42 (1), 247-255.
- Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56 (1), 137-172.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.
- Cross, L.H. Impara, J.C., Frary, R.B., & Jaeger, R.M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.
- Ebel, R.L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.
- Educational Testing Service (ETS) (1976). Report on a study of the use of the National Teachers' Examination by the state of South Carolina. Princeton, NJ: Author.
- Hambleton, R.K. & Eignor, D.R. (1980). Competency test development, validation, and standard setting. In R.M. Jaeger, & C.K. Tittle, (Eds.), Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 367-396). Berkeley, CA: McCutchan.

Table 1

Summary Statistics By Judging Method and Data Condition

<u>Method/Condition</u>	<u>n</u>	<u>Mean</u>	<u>SD</u>	<u>Within Judge Item Variability</u>	<u>Inter-Judge Reliability</u>		<u>Judge Internal Consistency</u>	<u>Item Internal Consistency</u>
					<u>Total</u>	<u>Single</u>		
					<u>Judge</u>	<u>Judge</u>		
Angoff								
With Data	20	12.79	2.45	.18	.94	.43	.97	.96
Without Data	19	11.22	2.40	.20	.74	.13	.81	.88
Modified Angoff								
With Data	19	13.13	1.07	.20	.96	.57	.97	.73
Without Data	19	11.16	1.86	.23	.79	.16	.81	.75



GROUPED ITEM RESPONSE AND ITEM DIFFICULTY INDEX OVER 20 TEST ITEMS